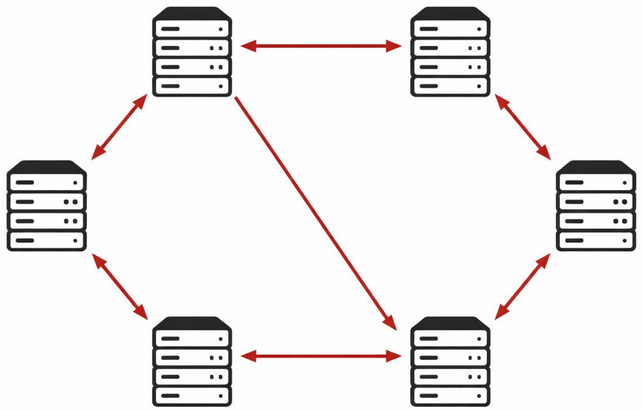
Accelerated Dual Method for Distributed Optimization: An Inexact-Gradient View of Local Updates

Junchi Yang, Ziyang Zeng, Linxuan Pan, Murat Yildirim, Feng Qiu

Presenter: Junchi Yang (Chinese University of Hong Kong, Shenzhen)

ICML 2026

Distributed Optimization



$$\min_{x \in \mathbb{R}^n} F(x) := \frac{1}{M} \sum_{i=1}^{M} f_i(x).$$

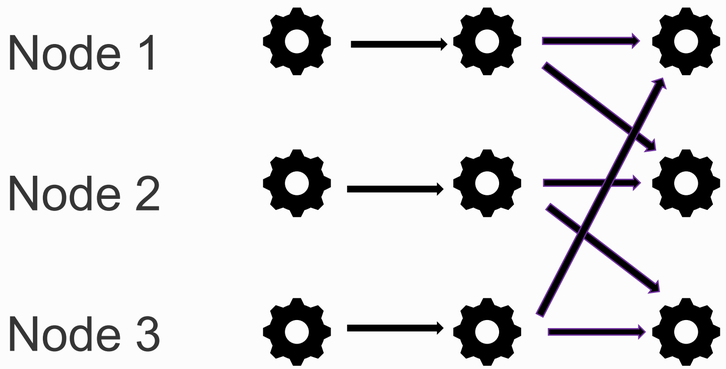
Local Updates for Distributed Optimization

Without Local Updates

Compute one local gradient/
One local update

Local Updates for Distributed Optimization

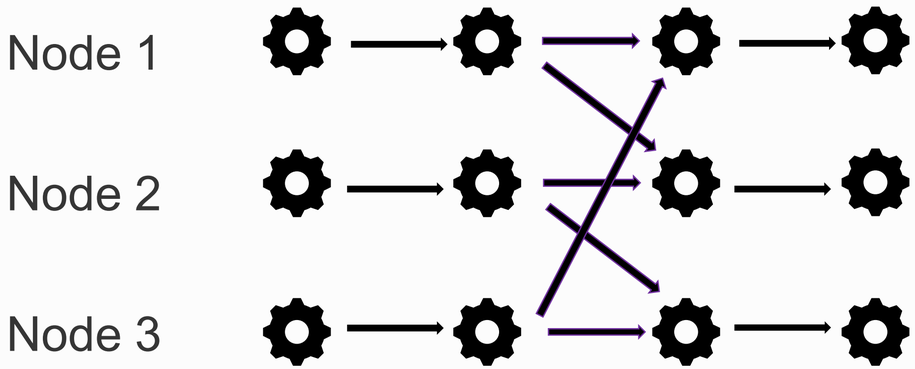
Without Local Updates



Communicate with neighbors

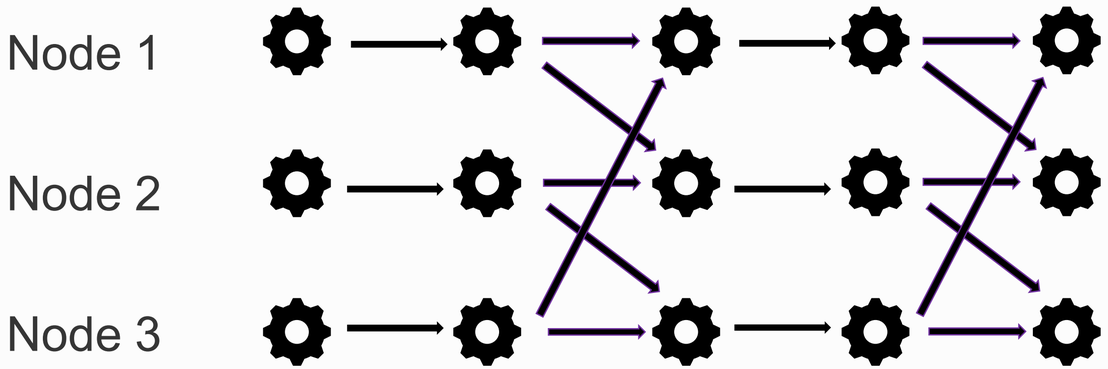
Local Updates for Distributed Optimization

Without Local Updates



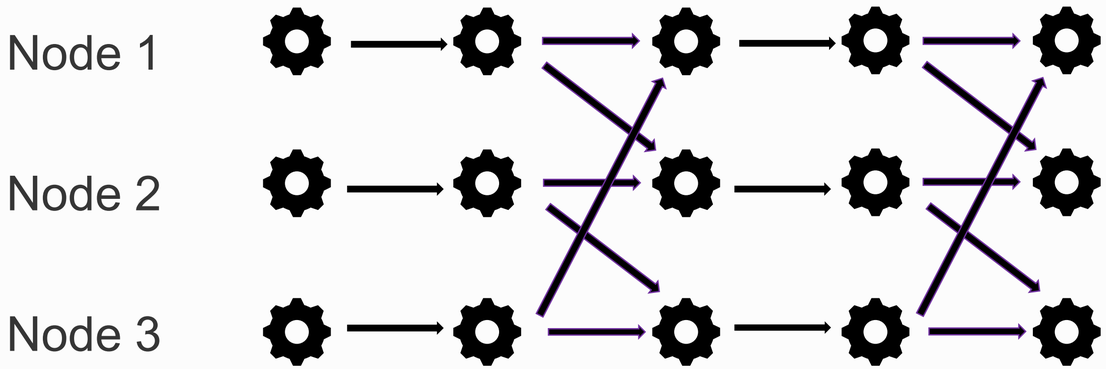
Local Updates for Distributed Optimization

Without Local Updates

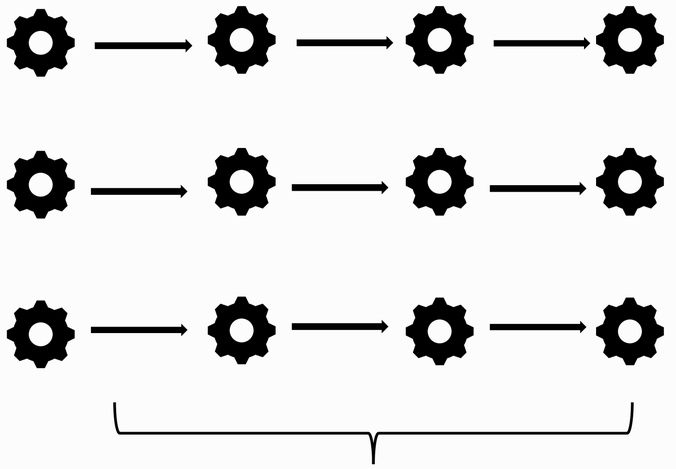


Local Updates for Distributed Optimization

Without Local Updates



With Local Updates



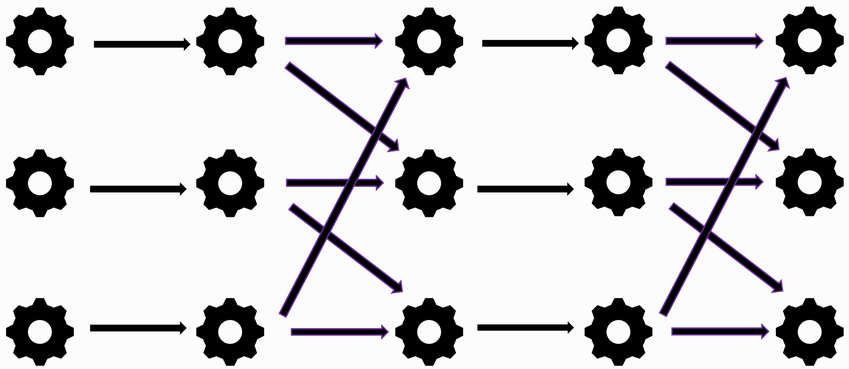
Local Updates for Distributed Optimization

Without Local Updates

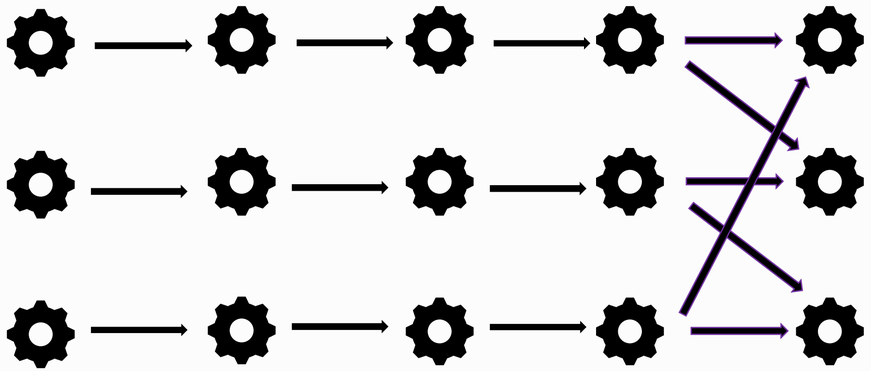
Node 1

Node 2

Node 3



With Local Updates



Communicate with neighbors

Strongly Convex Stochastic Distributed Optimization

- **Setting**: μ-strong convexity + ℓ-Lipschitz smoothness

  Condition number $\kappa = \ell/\mu$.

- **The mixing matrix W**: doubly stochastic, symmetric, and graph connected.

  Denote $p = 1 - \sigma_2$, where σ_2 is the second largest singular value of W.

Accelerated Dual Methods

- (Accelerated) dual methods have been widely studied [Terelius et al., 2011; Ghadimi et al., 2011; Scaman et al., 2017; Uribe et al., 2020].
- Define $U = (I - W)^{\frac{1}{2}}$. The original problem can be reformulated as

$$\min_{X \in \mathbb{R}^{M \times n}} \quad H(X) \qquad \text{where } X = (x^1, \ldots, x^M)^\top \in \mathbb{R}^{M \times n}$$

$$\text{s.t.} \quad UX = 0 \qquad H(X) = \frac{1}{M} \sum_{i=1}^{M} f_i(x^i)$$

- Lagrangian for this constrained problem:

$$\min_{X \in \mathbb{R}^{M \times n}} \max_{\lambda \in \mathbb{R}^{M \times n}} \mathcal{L}(X, \lambda) = H(X) + \langle \lambda, UX \rangle.$$

Accelerated Dual Methods

- Apply K steps of stochastic gradient descent on X and accelerated gradient ascent in λ

$$\textbf{Primal:} \quad \text{for } k = 0, \dots, K-1$$
$$X_{t,k+1} = X_{t,k} - \tau_2 \left(\widetilde{\nabla} H(X_{t,k}) + U \tilde{\lambda}_t \right)$$

$$\textbf{Dual:} \quad \lambda_{t+1} = \tilde{\lambda}_t + \tau_1 \, U X_{t,K}$$
$$\tilde{\lambda}_{t+1} = \lambda_{t+1} + \beta \left(\lambda_{t+1} - \lambda_t \right)$$

No communication
K Local updates

- Change of variable $\zeta = U\lambda$,

$$\textbf{Primal:} \quad \text{for } k = 0, \dots, K-1$$
$$X_{t,k+1} = X_{t,k} - \tau_2 \left(\widetilde{\nabla} H(X_{t,k}) + \tilde{\zeta}_t \right)$$

$$\textbf{Dual:} \quad \zeta_{t+1} = \tilde{\zeta}_t + \tau_1 \, W X_{t,K}$$
$$\tilde{\zeta}_{t+1} = \zeta_{t+1} + \beta \left(\zeta_{t+1} - \zeta_t \right)$$

$$\textbf{Primal (agent } i\textbf{):} \quad \text{for } k = 0, \dots, K-1$$
$$x_{i,t,k+1} = x_{i,t,k} - \tau_2 \left(\widetilde{\nabla} f_i(x_{i,t,k}) + \tilde{\zeta}_{i,t} \right)$$

$$\textbf{Dual (agent } i\textbf{):} \quad \zeta_{i,t+1} = \tilde{\zeta}_{i,t} + \tau_1 \sum_{j=1}^{M} w_{ij} \, x_{j,t,K}$$
$$\tilde{\zeta}_{i,t+1} = \zeta_{i,t+1} + \beta \left(\zeta_{i,t+1} - \zeta_{i,t} \right)$$

Accelerated Dual Methods

Proposition

The dual function $\Psi(\lambda) = \min_X \mathcal{L}(X, \lambda)$ is μ_Ψ-strongly concave and ℓ_Ψ-smooth in $Span(U)$, where $\mu_\Psi = Mp/\ell$ and $\ell_\Psi = 2M/\mu$.

- The condition number of Ψ is $2\kappa/p$. Applying Accelerated Gradient Ascent (AGA) to Ψ has the communication complexity of $\tilde{O}\left(\sqrt{\frac{\kappa}{p}}\right)$.

- This optimal co

? Does it converge under arbitrary K?

- When K is large

? Can it attain optimal communication complexity with $K = O(\epsilon^{-1})$?

Convergence of Accelerated Dual Methods

Theorem (Accelerated Dual Method)

With proper stepsize and momentum parameter, if the local steps

1. $K \leq \widetilde{\Theta}\left(\max\left\{\frac{c_0\sigma^2}{\epsilon}, \kappa\right\}\right)$, then the number of communication rounds to achieve an ϵ-accurate solution is $\tilde{O}\left(\frac{c_1}{\epsilon K}\right)$.

2. $K \geq \widetilde{\Theta}\left(\max\left\{\frac{c_0\sigma^2}{\epsilon}, \kappa\right\}\right)$, then the number of communication rounds to achieve an ϵ-accurate solution is $\tilde{O}\left(\left(\frac{\kappa}{p}\right)^{\alpha}\right)$ with $\alpha \in \left[\frac{1}{2}, 1\right]$, given $\beta = 1 - O\left((\mu\tau_1)^{1-\alpha}\right)$.

- When local updates $K \leq \widetilde{\Theta}\left(\max\left\{\frac{c_0\sigma^2}{\epsilon}, \kappa\right\}\right)$, the total sample complexity is $\tilde{O}\left(\frac{c_1}{\epsilon}\right)$.
- When local updates $K \geq \widetilde{\Theta}\left(\max\left\{\frac{c_0\sigma^2}{\epsilon}, \kappa\right\}\right)$, the communication complexity can be $\tilde{O}\left(\sqrt{\frac{\kappa}{p}}\right)$.

Summary for Strongly Convex Setting

Communication Complexities
when the number local steps/minibatch is large enough

	Communication Complexity	Local Updates	Arbitrary K
Local-DSGD (Koloskova et al., 2020)	$\widetilde{\mathrm{O}}\left(\kappa p^{-1} + \sqrt{\kappa} p^{-1} \epsilon^{-\frac{1}{2}}\right)$	Yes	Yes
LED (Alghunaim, 2024)	$\widetilde{\mathrm{O}}(\kappa^2 p^{-1})$	Yes	Yes
Stochastic GT (Koloskova et al., 2021)	$\widetilde{\mathrm{O}}(\kappa p^{-1} c^{-1})$	No	Yes
Distributed FGM (Uribe et al., 2020)	$\widetilde{\mathrm{O}}\left(\kappa^{\frac{1}{2}} p^{-\frac{1}{2}}\right)$	Yes	No
Local ADA (our work)	$\widetilde{\mathrm{O}}\left(\kappa^{\frac{1}{2}} p^{-\frac{1}{2}}\right)$	Yes	Yes
Lower Bound (Scaman et al., 2017)	$\widetilde{\Omega}\left(\kappa^{\frac{1}{2}} p^{-\frac{1}{2}}\right)$	N/A	N/A